Instantaneous Fundamental Frequency Estimation with Optimal Segmentation for Nonstationary Voiced Speech

Nørholm, Sidsel Marie; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2016.2608948

Publication date:
2016

Document Version
Early version, also known as pre-print

Instantaneous Fundamental Frequency Estimation with Optimal Segmentation for Non-Stationary Voiced Speech

Sidsel Marie Nørholm, Jesper Rindom Jensen, *Member, IEEE,*
and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—In speech processing, the speech is often considered stationary within segments of 20–30 ms even though it is well known not to be true. In this paper, we take the non-stationarity of voiced speech into account by using a linear chirp model to describe the speech signal. We propose a maximum likelihood estimator of the fundamental frequency and chirp rate of this model, and show that it reaches the Cramer-Rao lower bound. Since the speech varies over time, a fixed segment length is not optimal, and we propose making a segmentation of the signal based on the maximum a posteriori (MAP) criterion. Using this segmentation method, the segments are on average longer for the chirp model compared to the traditional harmonic model. For the signal under test, the average segment length is 24.4 ms and 17.1 ms for the chirp model and traditional harmonic model, respectively. This suggests a better fit of the chirp model than the harmonic model to the speech signal. The methods are based on an assumption of white Gaussian noise, and, therefore, two prewhitening filters are also proposed.**

***Index Terms*—Harmonic chirp model, parameter estimation, segmentation, prewhitening.**

I. Introduction

PARAMETER estimation of harmonic signals is relevant to the fields of speech processing and communication. In speech models, the speech signal is often split into a voiced part and an unvoiced part. The voiced part of the speech signal is produced by the vibration of the vocal cords, and, therefore, has a structure with a fundamental frequency and a set of overtones given by integer multiples of the fundamental. Over the years, several fundamental frequency estimators have been proposed based on different methods, such as autocorrelation [2], statistical [3]–[5], optimal filtering [6], or eigenvalue decomposition [7], [8]. Some methods work directly in the time domain [8], [9] whereas others use the spectrum or cepstrum [10], [11]. Comparisons of various fundamental frequency estimators have shown that different domains offer different advantages in e.g., the two genders [12]. Most of these fundamental frequency estimators split the signal into segments of 20–30 ms [13], make a voiced/unvoiced decision [14], [15], and estimate the parameters of each voiced segment separately. In most models, the signal is assumed stationary within each segment, even though it is well known that this assumption of stationarity does not hold [13], [16]. Some estimators overcome this problem of non-stationarity by looking at shorter segments, as, e.g., in [17], [18] where the fundamental frequency is estimated based on a single period of voiced speech. This overcomes the problem of non-stationarity, however, the lack of data points, that each estimate is based on, gives a greater uncertainty of the estimates. This is also seen in [18] where the method has a poor performance with respect to fine pitch error (FPE). Another approach, giving higher estimation accuracy, is to model the change in fundamental frequency within each segment. This can be done by extending the harmonic model [19]–[22] to a harmonic chirp model, which has also been suggested in [24], [25], [37]. Here, the harmonic structure remains the foundation of the model, but the fundamental frequency is allowed to change linearly within each segment. This introduces an extra parameter to estimate, but with the benefit that the model fits the speech signal better. Using the harmonic chirp model instead of the traditional harmonic model can, therefore, lead to better speech enhancement [26], but with a better fit of the model it is also possible to work with longer segments. In general, longer segments lead to better performance of the estimators, and so a smaller error on the estimated parameters can be obtained. However, the optimal segment length depends on the features of the signal, which are varying over time in the case of speech signals. At some time instances, the parameters are almost constant, and, in such periods, long segments can be used whereas at other points in time, the parameters will change fast and shorter segments should be used. Instead of using a fixed segment length, it is, therefore, better to have a varying segment length that depends on the signal characteristics at the given point in time. In [27], [28], the signal is modelled based on linear prediction (LP), and the segment length is chosen according to a trade-off between bit rate and distortion. The principle can, however, be used with other criteria for choosing the segment length, depending on what is most relevant in the given situation. The noise characteristics also have an impact on the performance of parameter estimators and optimal segmentation. Most methods make an assumption of white Gaussian noise, which is rarely experienced in real life scenarios. One way to address this problem is to preprocess the signal in a way that makes the noise resemble white Gaussian noise, as is, e.g., done through Cholesky factorisation [29].

The contribution in this paper is three-fold. First, we pro-

This work was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084. Part of this material was published at Asilomar 2014 [1].

S. M. Nørholm, J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, AD:MT, Aalborg University, DK-9000 Aalborg, Denmark, e-mail: {smn, jrj, mgc}@create.aau.dk

pose estimating the fundamental frequency and fundamental chirp rate by maximising the likelihood. Since maximising with respect to two parameters leads to a search in a two-dimensional space, we suggest an iterative procedure where first a one dimensional optimisation of the chirp parameter is performed followed by a one dimensional optimisation of the fundamental frequency based on the newly found estimate of the chirp rate. The estimation process is ended by convergence of the two-dimensional cost function. The proposed parameter estimator is a continuation of [1]. Our iterative procedure offers some benefits over the method suggested in [25], where an approximate cost function is introduced in order to decrease the computational load. The approximate cost function in [25] is evaluated over a two-dimensional grid, which means that fundamental frequency and chirp rate have to be found for each point in the grid before the optimum is found. In this paper, the original cost function is evaluated iteratively, giving fewer points for evaluation thus making the procedure suggested in this paper faster. Second, we suggest a maximum a posteriori (MAP) criterion to either make model selection between the traditional harmonic model and the harmonic chirp model, or make optimal segmentation of the signal based on one of the models. The optimal segmentation is based on the principle suggested in [27], [28]. The principle is adapted to the harmonic chirp model by using the maximum a posteriori (MAP) criterion for choosing the segment length. The model selection and optimal segmentation are introduced to give better representations of the signal. With the model selection, the more complex harmonic chirp model is favoured over the traditional harmonic model whenever it is beneficial according to the MAP principle. This reduces the error in, e.g., reconstruction or filtering [26] of the signal while keeping complexity low by choosing the traditional harmonic model whenever this is sufficient. With optimal segmentation, the segment length differs over time, optimising the fit of the model to the signal in each segment. This results in parameters that better describe the signal in the segment, and so also a lower error on, e.g., reconstruction or filtering. Third, we suggest two different methods to prewhiten the noise. Both the maximum likelihood estimator of the fundamental frequency and chirp rate and the MAP criterion are based on an assumption of white Gaussian noise, and, therefore, a prewhitening step is necessary if the noise is not white Gaussian. Both methods are based on noise power spectral density (PSD) estimation [30]–[33] and generate a filter to counteract the spectral shape of the noise. The filter is either based directly on the estimated spectrum of the noise or linear prediction of the noise.

The paper is organised as follows. In Section II, the harmonic chirp model is introduced. In Section III, the maximum likelihood estimator of the fundamental frequency and fundamental chirp rate is derived. In Section IV, the general MAP criterion is introduced for the harmonic chirp model along with the MAP model selection criterion between the traditional harmonic model, the harmonic chirp model and the noise only model. This is followed by the segmentation principle based on the MAP criterion in Section V. In Section VI, the two prewhitening methods are described. In Section VII, the proposed methods are tested through simulations on synthetic chirp signals and speech, and the paper is concluded in Section VIII.

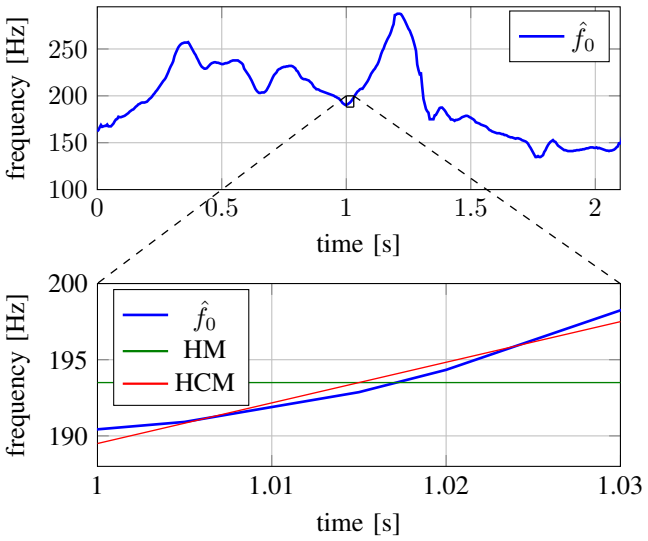


Fig. 1: Sketch of the difference between the harmonic model (HM) and the harmonic chirp model (HCM). The top plot shows a fundamental frequency track ($\hat{f}_0$) of a speech signal. The bottom plot is an enlargement of the small black square in the top plot.

II. Harmonic chirp model

In order to illustrate the difference between the harmonic model and the harmonic chirp model, a fundamental frequency track of a speech signal is plotted in the top of Fig. 1. The figure shows that the fundamental frequency changes continuously over time. This is also illustrated in the bottom figure with an enlargement of the 30 ms segment marked by the black square in the top figure. In this 30 ms segment, the fundamental frequency changes by approximately 8 Hz, whereas the harmonic model (HM), and most other fundamental frequency estimators, would assume the instantaneous fundamental frequency to be stationary within the segment. The harmonic chirp model (HCM) does not assume stationarity, but assumes a linear change of the fundamental frequency within a segment. As shown in the bottom figure, this model better describes the instantaneous fundamental frequency in the segment. With a better model, it is possible to work with longer segments, which will give higher accuracy on the estimated parameters. Further, it can lead to more efficient coding and signal reconstruction.

The harmonic chirp model is an extension of the traditional harmonic model. Therefore, the frequencies of the harmonics are still given by integer multiples of a fundamental frequency. However, in the chirp model, the instantaneous frequency of the l'th harmonic, $\omega_l(n)$, varies with the time index $n = n_0, ..., n_0 + N - 1$ in a linear way:

$$\omega_l(n) = l(\omega_0 + kn), \tag{1}$$

where $\omega_0 = 2\pi f_0/f_s$, with f_s the sampling frequency, is the normalised fundamental frequency, and k is the normalised

fundamental chirp rate. This means that in order to obtain the instantaneous frequency, both the fundamental frequency and the chirp rate are needed. The instantaneous phase, $\varphi_l(n)$, of the sinusoids are given by the integral of the instantaneous frequency as

$$\varphi_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l, \tag{2}$$

where $\phi_l \in [0, 2\pi]$ is the initial phase of the l'th harmonic. This leads to the complex harmonic chirp model for a voiced speech signal, $s(n)$:

$$s(n) = \sum_{l=1}^{L} A_l e^{j\varphi_l(n)} \tag{3}$$

$$= \sum_{l=1}^{L} \alpha_l e^{jl\left(\omega_0 n + k/2n^2\right)}, \tag{4}$$

where L is the number of harmonics and $\alpha_l = A_l e^{j\phi_l}$, $A_l > 0$ is the complex amplitude of the l'th harmonic. For speech signals the model order has to be estimated, which can be done, e.g., by use of the MAP criterion introduced in Section IV (see also [8]). The complex signal model is used instead of the real because it can ease both notation and computation. A real signal can be easily converted to a complex signal by use of the Hilbert transform [34] and without loss of information, downsampled by a factor of two.

A special case of the harmonic chirp model for $k = 0$ is the traditional harmonic model:

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n}. \tag{5}$$

Defining a vector of samples

$$\mathbf{s} = [s(n_0)\ s(n_0+1)\ \ldots\ s(n_0+N-1)]^T, \tag{6}$$

where $(\cdot)^T$ denotes the transpose. Note that the dependency on the index n_0 is left out for ease of notation. The signal model is then written as

$$\mathbf{s} = \mathbf{Z}\mathbf{a}, \tag{7}$$

where $\mathbf{Z}$ is a matrix constructed from a set of L modified Fourier vectors matching the harmonics of the signal,

$$\mathbf{Z} = [\mathbf{z}(\omega_0, k)\ \ \mathbf{z}(2\omega_0, 2k)\ \ \ldots\ \ \mathbf{z}(L\omega_0, Lk)], \tag{8}$$

with

$$\mathbf{z}(l\omega_0, lk) = \begin{bmatrix} e^{jl(\omega_0 n_0 + k/2n_0^2)} \\ e^{jl(\omega_0(n_0+1)+k/2(n_0+1)^2)} \\ \vdots \\ e^{jl(\omega_0(n_0+N-1)+k/2(n_0+N-1)^2)} \end{bmatrix}. \tag{9}$$

The vector $\mathbf{a}$ contains the complex amplitudes of the harmonics, $\mathbf{a} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_L]^T$.

The signal we want to make parameter estimation on, is often buried in noise, $v(n)$, to give the observed signal, $x(n)$,

$$x(n) = s(n) + v(n), \tag{10}$$

which can also be put into a vector of observed samples

$$\mathbf{x} = \mathbf{s} + \mathbf{v}, \tag{11}$$

where $\mathbf{x}$ and $\mathbf{v}$ are defined similarly to $\mathbf{s}$ in (6). For real signals as speech, the signal model will not fit the desired signal perfectly, and so $\mathbf{v}$ will also cover the part of the speech signal that does not align with the given model as, e.g., unvoiced speech during mixed excitations.

III. Estimation of frequency and chirp rate

The fundamental frequency and chirp rate are estimated by maximising the likelihood. The maximum likelihood estimates are the parameters of the model that describe the observed signal the best, i.e., the parameters that maximise the probability of the observed data, $\mathbf{x}$, given the parameters:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}), \tag{12}$$

where $\boldsymbol{\theta}$ is a vector containing the parameters of the model. Under the assumption of circularly symmetric Gaussian noise, the likelihood function can be written as [8]:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\pi^N \det(\mathbf{R}_v)} e^{-(\mathbf{x}-\mathbf{s})^H \mathbf{R}_v^{-1}(\mathbf{x}-\mathbf{s})} \tag{13}$$

$$= \frac{1}{\pi^N \det(\mathbf{R}_v)} e^{-\mathbf{v}^H \mathbf{R}_v^{-1}\mathbf{v}}, \tag{14}$$

where $\det(\cdot)$ denotes the determinant of the argument, $(\cdot)^H$ the Hermitian transpose and $\mathbf{R}_v = \mathbb{E}[\mathbf{v}\mathbf{v}^H]$ the noise covariance matrix, with $\mathbb{E}(\cdot)$ the mathematical expectation. Often the log likelihood is maximised instead of the likelihood

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - \ln \det(\mathbf{R}_v) - \mathbf{v}^H \mathbf{R}_v^{-1}\mathbf{v}. \tag{15}$$

In the case of white noise, the noise covariance matrix reduces to a diagonal matrix, $\mathbf{R}_v = \sigma_v^2 \mathbf{I}_N$, where σ_v^2 is the variance of the noise signal and $\mathbf{I}_N$ is an $N \times N$ identity matrix. The log likelihood can, therefore, be reduced to

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - N \ln \sigma_v^2 - \frac{1}{\sigma_v^2}||\mathbf{v}||_2^2. \tag{16}$$

The noise and its variance can be found using the signal model in (7)

$$\mathbf{v} = \mathbf{x} - \mathbf{s} = \mathbf{x} - \mathbf{Z}\mathbf{a} \Rightarrow \tag{17}$$

$$||\mathbf{v}||_2^2 = ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2, \tag{18}$$

$$\sigma_v^2 = \frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2, \tag{19}$$

which turns the log likelihood into

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = -N \ln \pi - N \ln \frac{1}{N}||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2 - N. \tag{20}$$

In the estimation of the fundamental frequency and chirp rate, it is only necessary to consider terms dependent on these two parameters, and the log likelihood function can be reduced to the nonlinear least squares (NLS) estimator that minimises the error between the observed signal and the signal model:

$$\{\widehat{\mathbf{a}}, \widehat{\omega}_0, \widehat{k}\} = \arg\min_{\mathbf{a},\omega_0,k} ||\mathbf{x} - \mathbf{s}||_2^2 \tag{21}$$

$$= \arg\min_{\mathbf{a},\omega_0,k} ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2. \tag{22}$$

Here, we are interested in the joint estimation of the fundamental frequency and chirp rate, and, therefore, the amplitudes are substituted with their least squares estimate [9],

$$\widehat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}, \tag{23}$$

to give the estimator:

$$\{\widehat{\omega}_0, \widehat{k}\} = \arg\min_{\omega_0,k} ||\mathbf{x} - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}||_2^2 \tag{24}$$

$$= \arg\min_{\omega_0,k} \left(\mathbf{x}^H(\mathbf{I}_N - \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H)\mathbf{x}\right) \tag{25}$$

$$= \arg\min_{\omega_0,k} \left(\mathbf{x}^H\Pi^{\perp}\left(\omega_0, k\right)\mathbf{x}\right), \tag{26}$$

where Π is an orthogonal projection matrix

$$\Pi\left(\omega_0, k\right) = \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H \tag{27}$$

and $\Pi^{\perp}$ its orthogonal complement

$$\Pi^{\perp}\left(\omega_0, k\right) = \mathbf{I}_N - \Pi\left(\omega_0, k\right). \tag{28}$$

This process includes a two-dimensional optimisation over ω_0 and k. To solve the problem in a computationally efficient manner, we propose iterating between two one-dimensional searches [1]. First, the chirp rate in step i, k^i, is estimated using the fundamental frequency estimate from the previous iteration, $\omega_0^{(i-1)}$, $i = 1, 2, ...$

$$k^{(i)} = \arg\min_{k} \left(\mathbf{x}^H\Pi^{\perp}(\omega_0^{(i-1)}, k)\mathbf{x}\right). \tag{29}$$

This estimate of the chirp rate is used to find a new estimate of the fundamental frequency

$$\omega_0^{(i)} = \arg\min_{\omega_0} \left(\mathbf{x}^H\Pi^{\perp}(\omega_0, k^{(i)})\mathbf{x}\right). \tag{30}$$

The estimates of ω_0 and k are found by iterating between (29) and (30) until convergence of the cost function in (26), but could alternatively be ended by the convergence of the estimated parameters. The fundamental frequency and chirp rate minimising the cost function in (26) are found by searching among candidates in a grid centred at the value of the parameter from the previous iteration, $i - 1$. The grid search is followed by a Dichotomous search [35] to get a refined estimate of the minimum. It is expected that the fundamental frequency estimate is close to the estimate found under the assumption of stationarity within the analysis frame. Therefore, a fundamental frequency estimate found under the traditional harmonic assumption, e.g., by using one of the methods in [8], will be a good choice as an initialisation of the iterations, i.e., $\omega_0^{(0)} = \omega_{0,h}$. The chirp rate is expected to be small and the first grid search is, therefore, centred around zero, i.e., $k^{(0)} = 0$. The estimation process is summarised in Table I.

The best obtainable performance of an unbiased estimator is given by the Cramer-Rao lower bound (CRLB). The CRLB sets a lower limit to the variance of the parameter estimate

$$\text{var}(\widehat{\theta}_g) \geq [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{gg}, \tag{31}$$

where θ_g is the g'th parameter of the parameter vector $\boldsymbol{\theta}$ of length G, $[\cdot]_{gg}$ denotes the matrix element of row g and column g, and $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix (FIM) [36] of size $G \times G$:

$$[\mathcal{I}(\boldsymbol{\theta})]_{gh} = -\mathbb{E}\left\{\frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial\theta_g\partial\theta_h}\right\}. \tag{32}$$

Under the assumptions of white Gaussian noise and a noise covariance matrix independent of the parameters, the FIM reduces to:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{2}{\sigma_v^2}\text{Re}\left\{\frac{\partial\mathbf{s}^H}{\partial\boldsymbol{\theta}}\frac{\partial\mathbf{s}}{\partial\boldsymbol{\theta}^T}\right\} \tag{33}$$

$$= \frac{2}{\sigma_v^2}\text{Re}\left\{\mathbf{D}^H(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})\right\} \tag{34}$$

with

$$\mathbf{D}(\boldsymbol{\theta}) = [\mathbf{d}(\omega_0)\,\mathbf{d}(k)\,\mathbf{d}(A_1)\,\mathbf{d}(\phi_1)\,\ldots\,\mathbf{d}(A_L)\,\mathbf{d}(\phi_L)], \tag{35}$$

$$\mathbf{d}(y) = \frac{\partial\mathbf{s}}{\partial y}. \tag{36}$$

For the signal model at hand, the elements of the $\mathbf{d}$ vectors are:

$$[\mathbf{d}(\omega_0)]_n = \sum_{l=1}^{L} jlnA_l e^{jl(\omega_0 n + k/2n^2) + j\phi_l}, \tag{37}$$

$$[\mathbf{d}(k)]_n = \sum_{l=1}^{L} \frac{1}{2}jln^2A_l e^{jl(\omega_0 n + k/2n^2) + j\phi_l}, \tag{38}$$

$$[\mathbf{d}(A_l)]_n = e^{jl(\omega_0 n + k/2n^2) + j\phi_l}, \tag{39}$$

$$[\mathbf{d}(\phi_l)]_n = jA_l e^{jl(\omega_0 n + k/2n^2) + j\phi_l}. \tag{40}$$

The CRLB depends on the choice of n_0. The best estimates are obtained if the segment is centred around $n = 0$ [37], and, therefore, n_0 should be chosen as $n_0 = -(N-1)/2$ for N odd and $n_0 = -N/2$ for N even. The CRLB also depends on the number of harmonics and the amplitude of the l'th harmonic A_l. The CRLB for a harmonic signal [8] decreases with $A_l^2 l^2$, which means that the more harmonics included in the estimate of fundamental frequency and chirp rate, the better the estimate.

TABLE I: Estimation of fundamental frequency and chirp rate.

for each sample
initialisation
$\omega_0^{(0)} = \omega_{0,h}$
$k^{(0)} = 0$
$\Delta k = 2\alpha_k/(K-1)$
$\Delta\omega = 2\alpha_\omega/(K-1)$
repeat
$K = \{k^{(i-1)} - \alpha_k, \Delta k, ..., k^{i-1} + \alpha_k\}$
$\Omega = \{\omega_0^{(i-1)} - \alpha_\omega, \Delta\omega, ..., \omega_0^{i-1} + \alpha_\omega\}$
$k^{(i)} = \arg\min_{k\in K}\left(\mathbf{x}^H\Pi^{\perp}(\omega_0^{(i-1)}, k)\mathbf{x}\right)$
$\omega_0^{(i)} = \arg\min_{\omega_0\in\Omega}\left(\mathbf{x}^H\Pi^{\perp}(\omega_0, k^{(i)})\mathbf{x}\right)$
until (convergence)

IV. MAP CRITERION AND MODEL SELECTION

Model selection and segmentation can be done with a maximum a posteriori (MAP) model selection criterion. The principle behind the MAP criterion is to choose the model, $\mathcal{M}$,

that maximises the posterior probability given the observed data, $\mathbf{x}$:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathcal{M}|\mathbf{x}). \tag{41}$$

Using Bayes' theorem [38] this can be rewritten as:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} \frac{p(\mathbf{x}|\mathcal{M})p(\mathcal{M})}{p(\mathbf{x})}. \tag{42}$$

Choosing the same prior probability, $p(\mathcal{M})$, for every model to avoid favouring any model beforehand, and noting that the probability of a given data vector, $p(\mathbf{x})$, is constant once it has been observed, the MAP estimate can be reduced to:

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathbf{x}|\mathcal{M}), \tag{43}$$

which is the likelihood of the observed data given the model. The likelihood is also dependent on other parameters like the fundamental frequency and the model order. As opposed to the maximum likelihood approach, these have to be integrated out in the Bayesian framework to give the marginal density of the data given the model [8]:

$$p(\mathbf{x}|\mathcal{M}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}. \tag{44}$$

An approximation to this integral can be found assuming high amounts of data and a likelihood that is highly peaked around the maximum likelihood estimates of $\boldsymbol{\theta}$ [8], [23], [39]

$$p(\mathbf{x}|\mathcal{M}) = \pi^{G/2} \det(\widehat{\mathbf{H}})^{-1/2} p(\mathbf{x}|\widehat{\boldsymbol{\theta}}, \mathcal{M}) p(\widehat{\boldsymbol{\theta}}|\mathcal{M}), \tag{45}$$

where $\widehat{\mathbf{H}}$ is the Hessian of the log-likelihood function evaluated at $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\mathbf{H}} = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{46}$$

Now an expression for the MAP estimator can be found by taking the negative logarithm of (45). The term $\pi^{G/2}$ can be assumed constant for large N and is, therefore, neglected, while a weak prior on $p(\boldsymbol{\theta}|\mathcal{M})$ has been used [23] to obtain the expression [8]:

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} -\ln \mathcal{L}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \frac{1}{2} \ln \det(\widehat{\mathbf{H}}). \tag{47}$$

This corresponds to minimising a cost function, where the first part is the likelihood from (16), and the second part is a model-dependent penalty term.

The penalty term is found by noting that the Hessian is related to the Fisher information matrix in (32). Evaluating the Fisher information matrix at $\theta = \widehat{\theta}$ gives the expected value of the Hessian, and, therefore, the elements in the Hessian can be found by using (35)-(40). To ease complexity, an asymptotic expression for the Hessian can be found by looking at the elements of the matrix. The diagonal elements of the Hessian for the harmonic chirp model are given by:

$$\widehat{\mathbf{H}}_{\omega_0\omega_0} = \sum_{l=1}^{L} \frac{1}{12}(N^3 - N) l^2 \widehat{A}_l^2, \tag{48}$$

$$\widehat{\mathbf{H}}_{kk} = \sum_{l=1}^{L} \frac{1}{960}(3N^5 - 10N^3 + 7N) l^2 \widehat{A}_l^2, \tag{49}$$

$$\widehat{\mathbf{H}}_{A_l A_l} = N, \tag{50}$$

$$\widehat{\mathbf{H}}_{\phi_l \phi_l} = N \widehat{A}_l^2, \tag{51}$$

for N odd and $n_0 = -(N-1)/2$. From this, when the Hessian is evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, the model order and amplitudes can be considered constant, and the Hessian is then only dependent on N. To make this dependency negligible, a diagonal normalisation matrix, $\mathbf{K}$, is introduced [8], [40]

$$\mathbf{K} = \begin{bmatrix} N^{-3/2} & & \mathbf{0} \\ & N^{-5/2} & \\ \mathbf{0} & & N^{-1/2}\mathbf{I}_{2L} \end{bmatrix}, \tag{52}$$

resulting in

$$\widehat{\mathbf{H}} = \mathbf{K}^{-1} \mathbf{K} \widehat{\mathbf{H}} \mathbf{K} \mathbf{K}^{-1}. \tag{53}$$

The definition of the elements in $\mathbf{K}$ as $N^{-x/2}$ instead of N^{-x}, where $x = 1, 3, 5$, and multiplication with $\mathbf{K}$ from both sides is done to ensure that also the off-diagonal elements of $\widehat{\mathbf{H}}$ are compensated for in the right way. The determinant of the Hessian is then given by:

$$\det(\widehat{\mathbf{H}}) = \det(\mathbf{K}^{-2}) \det(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}), \tag{54}$$

where the main dependency on N is now moved to the term $\mathbf{K}^{-2}$ whereas $\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}$ is assumed small and constant for large N. Taking the natural logarithm of the determinant gives:

$$\ln \det(\widehat{\mathbf{H}}) = \ln \det(\mathbf{K}^{-2}) + \ln \det(\mathbf{K}\widehat{\mathbf{H}}\mathbf{K}) \tag{55}$$

$$= 3 \ln N + 5 \ln N + 2L \ln N + \mathcal{O}(1). \tag{56}$$

An expression for the cost associated with the harmonic chirp model can now be found by combining the log likelihood for the harmonic chirp model in (20) with the penalty term in (56) where the term $\mathcal{O}(1)$ is ignored:

$$J_c = N \ln \pi + N \ln \frac{1}{N} ||\mathbf{x} - \mathbf{Z}\mathbf{a}||_2^2 + N + \frac{3}{2} \ln N + \frac{5}{2} \ln N + L \ln N. \tag{57}$$

For the traditional harmonic model, the Hessian will not contain a term related to the chirp rate, k, and the penalty for the MAP estimator will, therefore, also be short of this term:

$$J_h = N \ln \pi + N \ln \frac{1}{N} ||\mathbf{x} - \mathbf{Z_0}\mathbf{a}||_2^2 + N + \frac{3}{2} \ln N + L \ln N, \tag{58}$$

where $\mathbf{Z}_0$ equals $\mathbf{Z}$ for $k = 0$. The MAP expressions for the harmonic chirp model and the traditional harmonic model can be used to choose between them by choosing the one with the smallest cost. Due to Occam's razor [41], the simplest model is always preferred if the models describe the signal equally well.

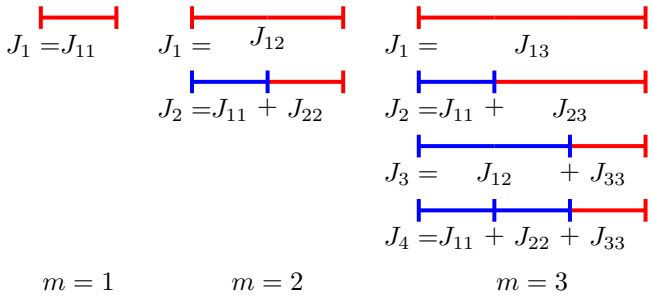


Fig. 2: Principle of segmentation. $M = 3$. Modified from [27].

This is assured by the extra penalty that naturally appears in the MAP expression for the chirp model. The error between the chirp model and the observed signal has to decrease enough relative to the traditional harmonic model to outweigh this penalty term before the chirp model is favoured over the traditional harmonic model. Aside from choosing between the two different harmonic models, the MAP estimator can also be used for voiced/unvoiced detection by determining whether a harmonic signal is present or not by comparing the two models with a zero order model,

$$J_0 = N \ln \pi + N \ln \sigma_x^2 + N, \tag{59}$$

where σ_x^2 is the variance of the observed signal. The voiced/unvoiced detection can also be done by using the generalised likelihood ratio test (GLRT) [42], [43]. In this method, the ratio of the likelihood of the presence of voiced speech found based on the harmonic model to the likelihood of a noise-only signal is calculated and compared to a threshold. The method has a constant false alarm ratio (CFAR) and so the threshold is set to ensure a given CFAR that is independent of the signal-to-noise ratio (SNR). Other methods as, e.g, described in [14], [15] can also be used.

V. Segmentation

The characteristics of the observed signal are varying over time and sometimes faster than others, meaning that a fixed segment length is not optimal. Using the MAP criterion, the cost associated with different segment lengths can be compared and the optimal chosen being the one minimising (57). The segmentation assures that the optimal trade-off between segment length and fit of the model is found, and so the segment length is chosen as long as possible without introducing too large modelling errors. It follows from the CRLB that long segments are desired and gives higher estimation accuracy. The segmentation is based on the principle in [27], [28] which is sketched in Fig. 2. In the figure, J_{xy} is the cost of a segment starting at block x and ending at block y, with both block x and y included in the segment.

A minimal segment length, N_{min}, is chosen, generating a block of N_{min} samples and dividing the signal into M blocks. Since this will give 2^{M-1} ways of segmenting the signal, a maximum number of blocks in one segment, K_{max}, is also set since very long segments are highly unlikely, and setting a maximum will bound the computational complexity. The maximum number of samples in one segment is, therefore, $N_{\text{max}} = K_{\text{max}} N_{\text{min}}$. Using a dynamic programming algorithm, the optimal number of blocks in a segment, k_{opt}, is found for all blocks, $m = 1, ..., M$, starting at $m = 1$ moving continuously to $m = M$. For each block, the cost of all new block combinations is calculated while old combinations are reused from earlier blocks. Relating to Fig. 2, the red segments are calculated whereas the blue segments are reused from earlier. To decrease the number of calculations further, only a block combination minimising the cost is used in a later step, which in Fig. 2 means that only one of J_3 and J_4 is considered for $m = 3$, corresponding to the block combination that minimised the cost at $m = 2$. When the end of the signal is reached, backtracking is used to find the optimal segmentation of the signal, starting at the last block, and jumping through the signal to the beginning. This is done by starting at $m = M$ and setting the number of blocks in the last segment of the signal to $k_{\text{opt}}(M)$. In this way, the next segment ends at block $m = M - k_{\text{opt}}(M)$ and includes $k_{\text{opt}}(M - k_{\text{opt}}(M))$ blocks. This is continued until $m = 0$. The segmentation is summarised in Table II.

TABLE II: Segmentation.

while $\boldsymbol{m \times N_{\text{min}} \leq}$ **length(signal)**
  $K = \min([m, K_{\text{max}}])$
  for $\boldsymbol{k = 1 : K}$
    blocks of signal to use is $m - k + 1, ..., m$
    find analytic signal and downsample
    estimate ω_0 and k using Table I
    estimate $\mathbf{a}$ and $\mathbf{Z}$ from (23), (8) and (9)
    calculate $J_{(m-k+1)m}$ from (57)
    $J(k) = \begin{cases} J_{(m-k+1)m} + J_{1(m-k)} & \text{if } m-k>0, \\ J_{(m-k+1)m} & \text{otherwise.} \end{cases}$
  end for
  $k_{\text{opt}}(m) = \arg\min J(k)$
  $m = m + 1$
end while

backtrack
$m = M$
while $\boldsymbol{m > 0}$
  number of blocks in segment is $k_{\text{opt}}(m)$
  $m = m - k_{\text{opt}}(m)$
end while

VI. Prewhitening

The maximum likelihood estimates of the fundamental frequency and chirp rate and the MAP model selection and segmentation criterion were found under the assumption of white Gaussian noise. However, in real life scenarios the noise is not always white. A prewhitening step is therefore required. The observed signal can be prewhitened by passing it through a filter that changes the noise from coloured to white. This is illustrated in Fig. 3. In the figure, $H(z)$ is a filter with a frequency response similar to the spectrum of the noise. The coloured noise can be seen as white noise filtered using a filter with coefficients given by $H(z)$. Therefore, to obtain a flat frequency spectrum of the noise, the action is reversed by dividing by $H(z)$, here denoted by $A(z)$. Naturally, the desired signal will also be altered by the passage through the filter. This may have an influence on the results depending on how much the signal is changed, and what the prewhitened signal

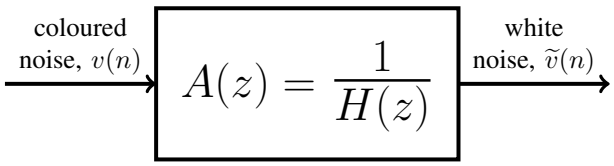


Fig. 3: Prewhitening of noise by passing it through the filter $A(z)$.

is used for. At the very best, the linear transformation of the signal will not affect the CRLB of the parameter estimation.

To obtain $H(z)$, information about the noise spectrum is needed. Different methods exist to estimate the power spectral density (PSD) of the noise given a mixture of desired signal and noise [30]–[33]. The PSD can be used directly to generate a simple finite impulse response (FIR) filter based on the frequency coefficients of the PSD. Alternatively, also based on the PSD, linear prediction (LP) can be used to find the characteristic parts of the noise spectrum and filter the observed signal based on this. In linear prediction, the present sample is estimated based on P prior samples:

$$\widehat{v}(n) = -\sum_{p=1}^{P} a_p v(n-p), \tag{60}$$

leading to a filter of the form:

$$A(z) = 1 + \sum_{p=1}^{P} a_p z^{-p}. \tag{61}$$

After filtering, the signal is normalised to have the same standard deviation before and after the filtering. To ensure that the desired signal has a smooth evolution over time after filtering, i.e., no drastic changes in amplitude or phase, it is important that the PSD is smooth. This is ensured by most recent PSD methods where the value in one time frame is a weighted combination of the preceding time frame and an estimate from the current time frame.

VII. Simulations

In the following, the different proposed methods are tested through simulations on synthetic signals and speech. The synthetic signals are made according to (7). Unless otherwise stated in the specific subsections, the signals were generated with $L = 10$, $A_l = 1 \forall l$, random phase, fundamental frequency, and fundamental chirp rate, in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [100, 300]$ Hz, $k \in [-500, 500]$ Hz/s and the sampling frequency, f_s, was set to 8000 Hz.

The speech signal, "Why were you away a year, Roy?", was used in some simulations and to illustrate the function of some methods. The sentence is uttered by a female speaker and sampled at 8000 Hz. Additionally, the five male and five female speech signals from the Keele database [44] are used. The signals have a duration of approximately 30 seconds each. The signals are downsampled to 8000 Hz. With these signals, follow the corresponding laryngograph signals and an annotated fundamental frequency that can be used for evaluation of the proposed method. However, it should be noted that the annotated fundamental frequency is also only an estimate and not the ground truth.

In most experiments, it is desirable to evaluate the methods at different SNRs, e.g., in an interval from -10 to 10 dB to simulate situations with different levels of background noise. Therefore, noise was added to the signals with a variance calculated to fit the desired input SNR defined as

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}, \tag{62}$$

where σ_s^2 is the variance of the desired signal. The noise signals used are white Gaussian noise, as well as different types of noise from the AURORA database [45].

For each segment of noisy speech, the discrete-time analytic signal [34] is computed, and the parameter estimation is performed on this complex, downsampled version of the signal.

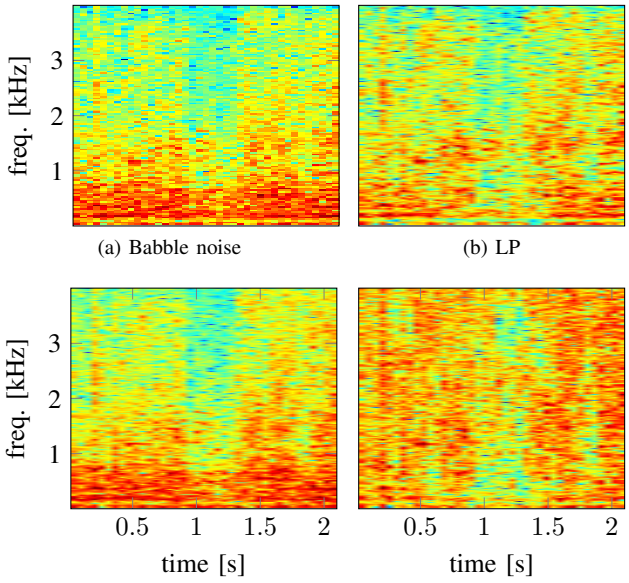
A. Prewhitening

The prewhitening using the FIR filter and LP is tested on "Why were you away a year, Roy?" and compared to prewhitening using Cholesky factorisation [46]. The signal is added noise at input SNRs of 0 and 10 dB, and the prewhitening filters are generated based on the noisy signal. The PSD is found using an implementation of [31] given in [30]. The PSD is obtained using 256 frequency points which equal the number of coefficients in the FIR filter, whereas the LP filter is made with five coefficients. The spectrum of babble noise at an input SNR of 10 dB before and after prewhitening is shown in Fig. 4. Here, it seems that the whitest noise signal is obtained using the Cholesky factorisation, followed by LP, while the FIR filter seems to make a minor change to the original noise.

The prewhitening methods are compared by means of the spectral flatness, $\mathcal{F}$, which is the ratio of the geometric mean to the arithmetic mean of the power spectrum, $S(k)$, [47]:

$$\mathcal{F} = \frac{\left(\prod_{k=0}^{K-1} S(k)\right)^{1/K}}{\frac{1}{K}\sum_{k=0}^{K-1} S(k)}. \tag{63}$$

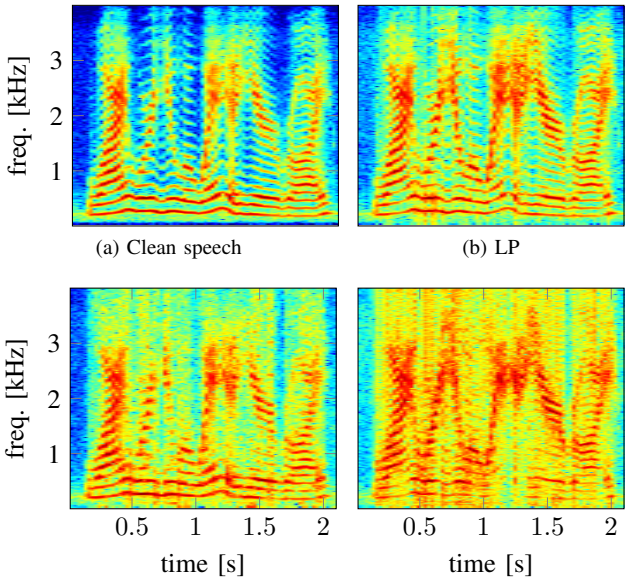
The spectral flatness gives a number between zero and one, where perfect white noise has a value of one. The spectral flatness for four different noise types at 0 and 10 dB is shown in Fig. 5, where the spectral flatness of the original noise and a white noise signal generated with MATLAB's `randn` are also shown for comparison. The spectral flatness is very similar at 0 and 10 dB for all noise types using a given prewhitening method. The results confirm the tendencies observed in Fig. 4. The Cholesky factorisation leads to the highest spectral flatness for all noise types, followed by linear prediction in the case of babble, car and street noise, while the FIR filter is better than linear prediction for exhibition noise. There is, however, large differences between the different noise types in how significant the advantage is of using one prewhitening method over another. The Cholesky factorisation is clearly best in terms of whitening the noise, but as is shown in Fig. 6, it is also the method that has the largest influence on the



(a) Babble noise (b) LP

(c) FIR (d) Cholesky

Fig. 4: Spectrograms of babble noise before (a) and after prewhitening with (b) LP filter, (c) FIR filter and (d) Cholesky factorisation. The four spectrograms are plotted with the same limits in dB.



(a) Clean speech (b) LP

(c) FIR (d) Cholesky

Fig. 6: Spectrograms of speech signal before (a) and after prewhitening with (b) LP filter, (c) FIR filter and (d) Cholesky factorisation. The four spectrograms are plotted with the same limits in dB.

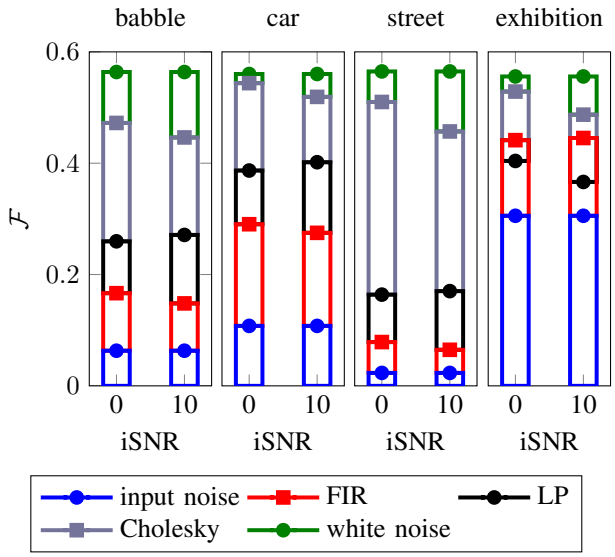


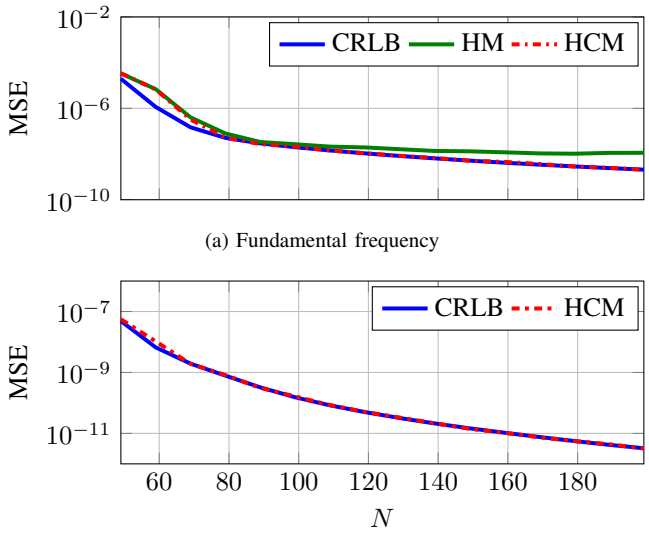
Fig. 5: Spectral flatness, $\mathcal{F}$, at 0 and 10 dB input SNR for original noise, prewhitened noise using FIR, LP and Cholesky factorisation. The spectral flatness for white noise is added for comparison.

desired signal. Here, it appears the LP filtering best preserves the desired signal with the FIR filter nearly as good, whereas the Cholesky factorisation clearly changes the appearance of the desired signal. Using the Cholesky factorisation for prewhitening, the signal model must be redefined to include the Cholesky matrix, as was done in [5]. Thus, it cannot be applied directly with the proposed model, and has been excluded from the following simulations. The FIR and LP filters only change the amplitude and phase, and, therefore, they only change the complex amplitude vector $\mathbf{a}$.

B. Fundamental frequency and chirp rate

The proposed estimator of fundamental frequency and chirp rate is first evaluated on synthetic chirp signals. Two experiments were made. In the first, the segment length, N, was varied from 49 to 199 samples with a fixed input SNR of 10 dB. In the second, the input SNR was varied from -10 to 10 dB with a fixed segment length of 199 samples. For each generated signal, noise was added, and an initial fundamental frequency estimate was found using a harmonic NLS estimator [8] with lower and upper limits of the search interval of 80 and 320 Hz. The model order is assumed known, i.e., $L = 10$. From here, the fundamental frequency and chirp rate were estimated, and the squared error was found. This was repeated 2000 times and the mean was taken to give the mean squared error (MSE). In Figs. 7 and 8, the MSE as a function of N and the input SNR is shown and compared to the CRLB and estimates obtained using a harmonic NLS estimator [8]. The chirp estimates reach the CRLB around a segment length of 110 and at an input SNR of around -5 dB under the given settings. The harmonic estimates are close to reaching the bound as well, but as the CRLB decreases for higher segment lengths and input SNRs, the error on the harmonic estimates do not decrease with the same rate resulting in a gap between the CRLB and the estimates.

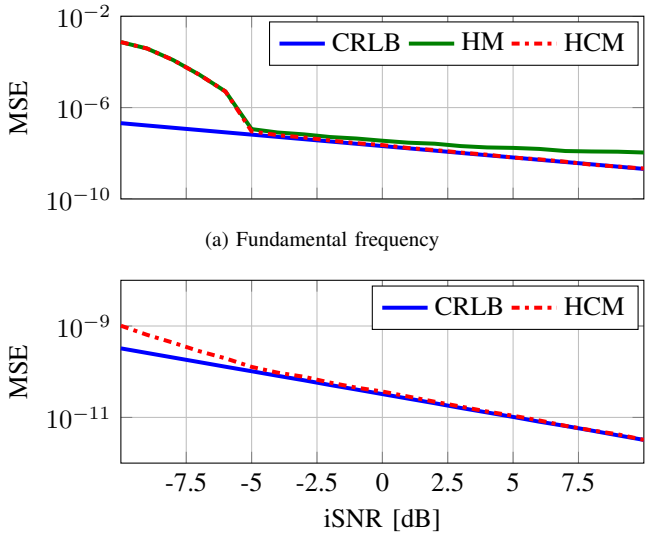
The estimator was used to estimate the fundamental frequency and chirp rate of "Why were you away a year, Roy?" with the spectrum shown in Fig. 6a. Here, the parameters are estimated directly from the clean signal in segments with



(a) Fundamental frequency

(b) Chirp rate

Fig. 7: Mean squared error (MSE) of the fundamental frequency and chirp rate as a function of N.



(a) Fundamental frequency

(b) Chirp rate

Fig. 8: Mean squared error for the fundamental frequency and chirp rate as a function of the input SNR.

a length of 198 samples (24.8 ms). The initial fundamental frequency estimate and model order were found jointly by using a harmonic NLS estimator and a MAP estimator [8], respectively. The limits on the harmonic fundamental frequency are set to 80 and 300 Hz. To confirm that the combination of the harmonic fundamental frequency and a chirp rate of zero is a good initialisation, an example of a two-dimensional cost function for a segment of a speech signal is shown in Fig. 9. The initialisation is marked by a yellow cross while the final estimate of fundamental frequency and chirp rate is marked by a red cross. As seen, the function is locally convex around the initial and true fundamental frequency and chirp rate. The figure also shows that the change in fundamental

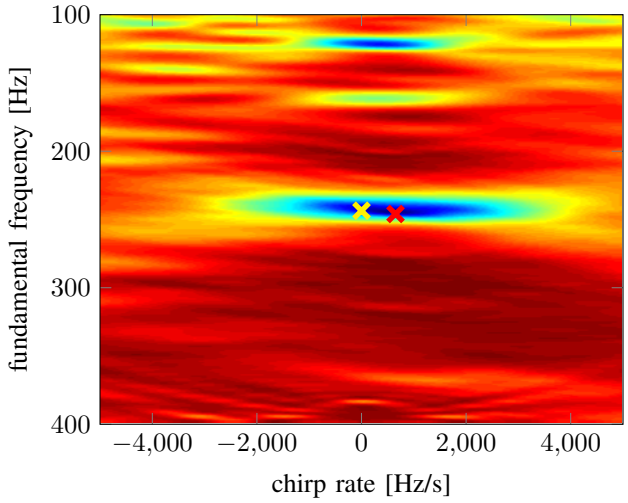


Fig. 9: Example of a cost function for a speech signal as a function of fundamental frequency and chirp rate.

frequency is rather small so if the fundamental frequency for some reason changes a lot, $\omega_0 < 0.6\omega_{0,\text{HM}}$ or $\omega_0 > 1.5\omega_{0,\text{HM}}$, the fundamental frequency is set to the harmonic estimate and the chirp rate is set to zero. However, it is important to note that the instantaneous fundamental frequency is not the same as the one found by the harmonic model. Now, the parameters are estimated in steps of 5 samples. The resulting estimates are shown in Fig. 10. The chirp rate can be interpreted as the tangent to the fundamental frequency curve at a given point. This means that the chirp rate should be negative when the fundamental frequency is decreasing, positive when it is increasing, and zero at a local maximum or minimum. To illustrate this, some maxima and minima of the fundamental frequency are marked by red stars in the upper plot and the chirp rates at the same points in time are marked in the bottom plot.

The estimation is repeated after the addition of noise to give an input SNR of 0 and 10 dB, but this time the parameters are only estimated once per segment of 198 samples. The estimation is done both for white Gaussian noise and babble noise as well as after prewhitening of the signal with babble noise using the FIR and LP filter. The sum of the absolute error between noisy and clean estimates is given in Table III at 0 and 10 dB. Here, only the time interval shown in Fig. 10 is considered since the beginning and end of the signal contain no speech. The white noise gives the best estimate at both 0 and 10 dB. At 0 dB, the LP prewhitened signal gives a lower error than the FIR filtered and clean babble noise whereas at 10 dB, the babble noise gives the lowest error followed by the FIR and LP filtered noise. This suggests that for the proposed ML estimator, the dominance of the desired signal at 10 dB decreases the importance of the noise shape relative to the effects of prewhitening on the signal. However, at 0 dB the noise is more dominant, and so the importance of prewhitening increases.

The fundamental frequency and chirp rate are also estimated from the signals in the Keele database. The fundamental

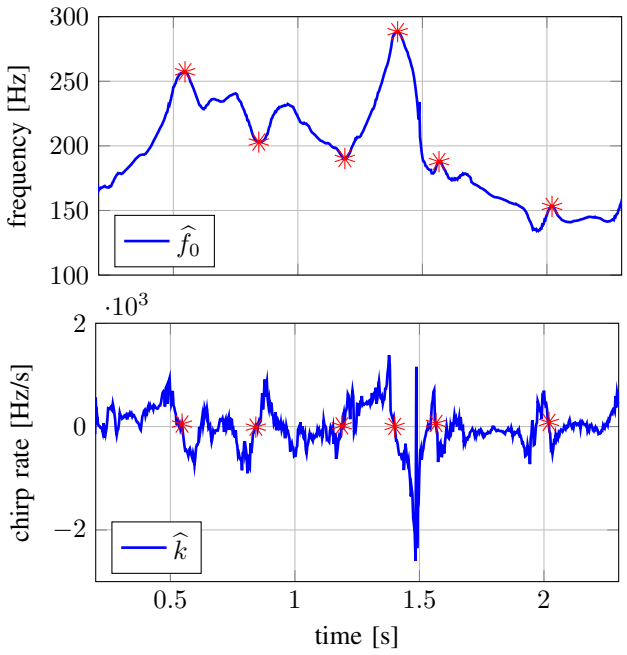


Fig. 10: Fundamental frequency and chirp rate estimation.

TABLE III: Sum of absolute error between noisy estimate and clean estimate of fundamental frequency in Hz at input SNRs of 0 and 10 dB.

	white noise	babble	FIR	LP
0 dB	585	2653	2483	1201
10 dB	167	408	714	787

frequency estimates are compared to YIN [2] and SWIPE [48] by means of the gross pitch error (GPE), the fine pitch error (FPE) and the reconstruction SNR. The GPE is defined as an estimate that deviates from the annotated fundamental frequency by more than 20 % [18]. The GPEs are not considered in the calculation of the FPE. The FPE is divided into two parts, the mean, μ, and the standard deviation, σ, of the errors on the estimates [12], [18]. Both are calculated from the difference between the estimated fundamental frequency and the annotated fundamental frequency. The annotated fundamental frequency is estimated in steps of 10 ms based on segments of 26.5 ms of data. This is also done for HM, HCM and YIN, however, it is not possible to choose the segment length in SWIPE. The lower and upper limit on the estimate are set to 50 and 300 Hz. The reconstruction SNR is calculated from the reconstructed signal based on (7). For YIN, SWIPE and the traditional harmonic model, the chirp rate is approximated by $\Delta f = (f_0(n+1) - f_0(n))/\Delta t$ where Δt is the time between two consecutive estimates of the fundamental frequency. Note that this will cause a delay in real-time applications. However, using past samples does not result in Δf for the correct segment and will degrade the reconstruction compared to only using the harmonic model. The estimated fundamental frequencies, chirp rates and Δf's are used in $\mathbf{Z}$ in (8). Since we are here considering non-stationary signals it makes a difference from where in the signal the reference point is set. From experiments on synthetic chirp signals it was found that YIN and SWIPE have the reference point towards the beginning of the signal whereas HM has its reference point around the middle. Therefore, we set $n_0 = 0$ for YIN and SWIPE and $n_0 = -N/2$ for HM. The mid-segment reference point for HM means that Δf is estimated incorrectly. The proper estimate would be $\Delta f = (f_0(n+1/2) - f_0(n-1/2))/\Delta t$, but this information is not available. The wrong estimate of Δf leads to a worse performance compared to using the harmonic model on its own. The result for HM without Δf is therefore also included in the comparison. The fundamental frequency, chirp rate and Δf are estimated for each 25 ms based on 25 ms of data, and the entire block of samples is reconstructed based on this estimate. The model order is estimated using a MAP estimator [8]. The amplitude vector, $\mathbf{a}$, is estimated using (23). The reconstruction SNR (rSNR) is then given by:

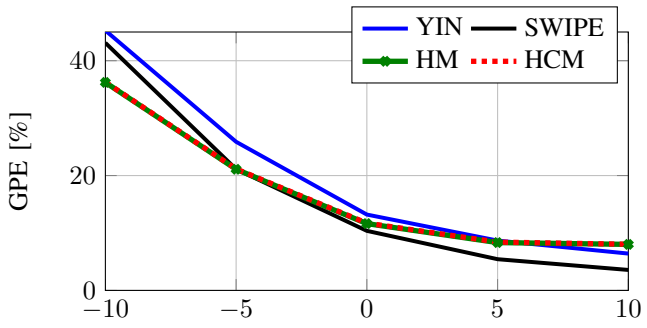
$$\text{rSNR} = \frac{\sigma_s^2}{\sigma_{(s-\hat{s})}^2}, \tag{64}$$

where $\hat{s}$ is the reconstructed signal, and $\sigma_{(s-\hat{s})}^2$ is the variance of the error signal between the original speech signal and the reconstructed signal.

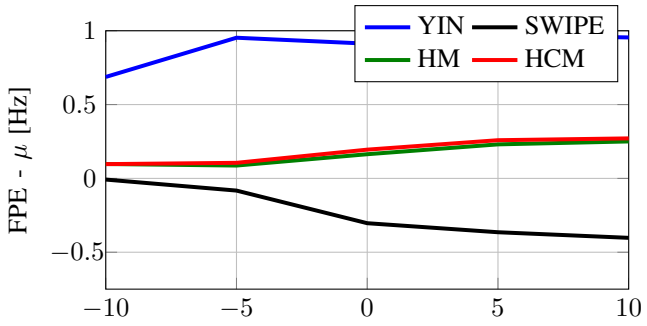
The results are shown in Fig. 11. In terms of GPE, the proposed method performs better than YIN and SWIPE at low input SNRs, while SWIPE is better at high input SNRs. The harmonic models perform equally. The bias, seen as the mean, μ, is small for all methods. It is approximately 1 Hz for YIN and within $\pm$ 0.5 Hz for the other methods. The proposed method does not perform as well as the traditional harmonic model in terms of standard deviation, σ. As mentioned earlier, the annotated fundamental frequency is not the ground truth, but a fundamental frequency estimate found from the laryngograph signal using an autocorrelation method which is also based on the harmonic assumption. In Fig. 9 it was seen that the instantaneous fundamental frequency found by the proposed method is not the same as the harmonic frequency. Therefore, it is not surprising that the method does not perform well when it is compared to the fundamental frequency estimated based on the harmonic assumption. Looking at the reconstruction SNR, the chirp model outperforms all other methods. The reconstruction SNR is the only of the four error measures that takes both fundamental frequency and chirp rate into account. Further, the reconstruction SNR does not depend on another estimate of the fundamental frequency as do the FPE and GPE, it compares to the original speech signal.

C. Model selection

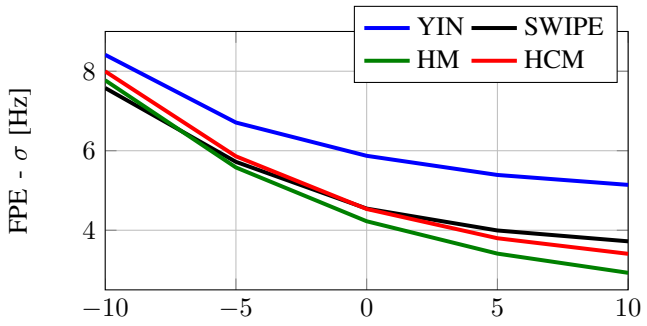
The model selection was first tested on synthetic signals degraded with white Gaussian noise to give an input SNR of 10 dB. In this part, the possible models included in the test are the traditional harmonic model and the harmonic chirp model. The model selection was tested for different chirp rates and different segment lengths. For each combination of chirp rate and segment length, 2000 signals were generated and the selected model was noted for each signal. The percent of the chirp model chosen is shown in Fig. 12. Even though



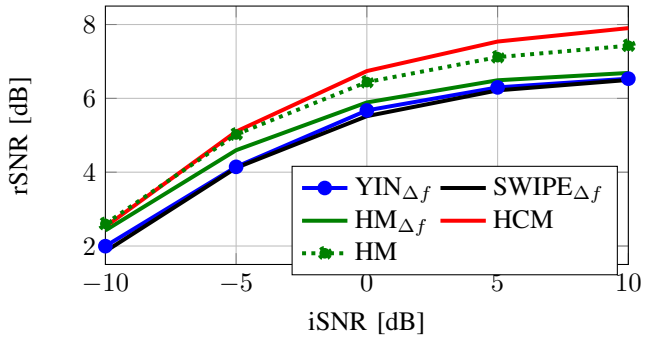
(a) Gross pitch error (GPE)



(b) Fine pitch error (FPE) - mean, μ



(c) Fine pitch error (FPE) - standard deviation, σ



(d) Reconstruction SNR

Fig. 11: Evaluation of the instantaneous fundamental frequency estimation by means of gross pitch error (GPE), fine pitch error (FPE) and reconstruction SNR (rSNR).

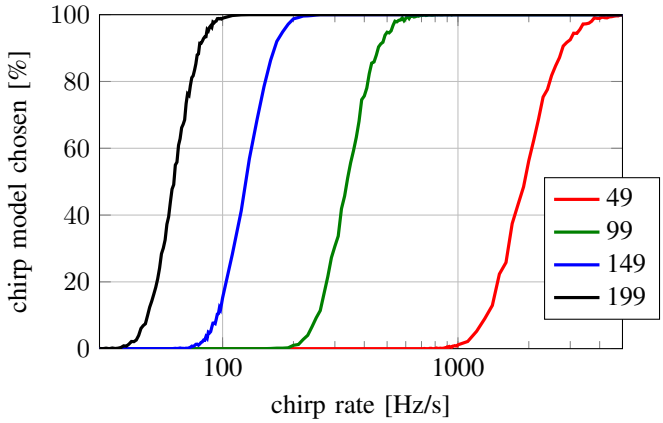


Fig. 12: Model selection for synthetic signals as a function of the chirp rate for different segment lengths from 49 to 199.

all generated signals, except for the ones with a chirp rate of zero, are chirp signals, the chirp model is not chosen in all cases. As mentioned in Section IV, this is due to the extra penalty term introduced to the chirp model and not to the harmonic model. The longer the signal is, the more prone it is to be denoted as a chirp signal since the error term $||\mathbf{x}-\mathbf{Za}||_2^2$ will increase with signal length when the model does not fit, making the cost of the harmonic model greater than that of the chirp model, despite the extra penalty to the chirp model.

Model selection was also performed on the speech signals from the Keele database in white Gaussian noise at different segment lengths. Here, the noise model is also included. The percentage of each chosen model is found by taking the number of segments labelled according to a given model out of the total number of segments in the signal. The result is shown in Fig. 13. The percentage of the chosen noise model is fairly independent of the segment length since the amount of unvoiced speech is independent of the segment length. For short segment lengths, the harmonic model is chosen approximately 55% of the time and the chirp model is never chosen, but as the segment length is increased, the two models are almost equally preferred. It should again be kept in mind that the chirp model has an extra penalty for being a more complex model so even though the error on the signal, $||\mathbf{x}-\mathbf{Za}||_2^2$, is smaller for the chirp model, it has to overcome the penalty as well before it is selected.

D. Segmentation

The segmentation is tested on the signal "Why were you away a year, Roy?". White Gaussian noise is added to the signal to give an input SNR of 10 dB. The signal is segmented according to the harmonic chirp model and the traditional harmonic model where, in both cases, the minimum segment length $N_{\text{min}} = 40$ and the maximum number of blocks $K_{\text{max}} = 10$, meaning that the minimum length of a segment is 40 samples (5 ms) and the maximum length of a segment, N_{max}, is 400 samples (50 ms). A representative example of the chosen segment length as a function of time is shown in Fig. 14. For comparison, the fundamental frequency estimate

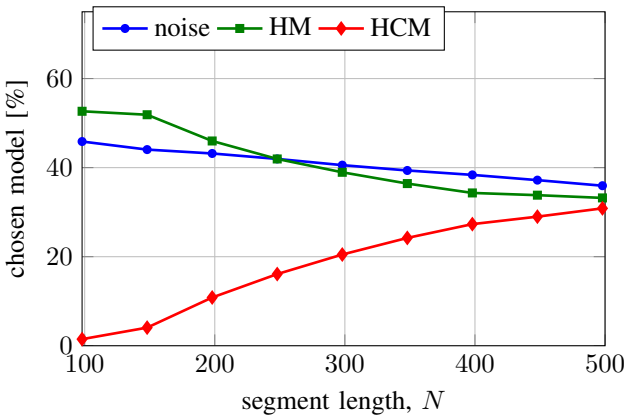


Fig. 13: Model selection as a function of the segment length (12.5 ms - 62.5 ms).

is plotted as well. In general, the chirp model gives rise to longer segment lengths than the traditional harmonic model. For this example, the average segment length is 195 samples (24.4 ms) using the chirp model and 137 samples (17.1 ms) using the traditional harmonic model. A typical choice of fixed segment length is 20–30 ms [13]. On average, this is a good choice when using the harmonic chirp model, however, shorter segments are better if the traditional harmonic model is used. The longer segments of the chirp model, of course, mean that the total number of segments is lower than for the harmonic model. The chirp model divides the signal into 105 segments and with the harmonic model, the number of segments is 150. Three areas in Fig. 14 are marked with circles as examples of the longer segments obtained with the chirp model. In the light blue circle, the fundamental frequency is decreasing quite fast, but the change is constant over time. Thus a long segment is obtained using the chirp model while shorter segments are obtained when the harmonic model is used. In the purple circle, the piece of speech is divided into four segments with the chirp model: two segments of maximum length, where the fundamental frequency is almost constant, and two shorter but still fairly long segments, where the fundamental frequency is increasing and decreasing, respectively. For the harmonic model, there are two long segments where the fundamental frequency is close to constant, but the rest of the piece is divided into shorter segments. In the brown circle, the piece is divided into two segments using the chirp model: one piece where the fundamental frequency is decreasing and one where it is increasing. The harmonic model covers the area in the middle, where the fundamental frequency is fairly constant, with two somewhat long segments, but in order to cover the whole area, shorter segments are added on both sides of the segments in the middle. The longer segments chosen for the chirp model suggests that the chirp model describes the signal in a better way than the traditional harmonic model since it to some extent takes the non-stationarity of the speech into account.

The signal is reconstructed using (7), as was done in the evaluation of the fundamental frequency estimate. The signal is reconstructed from the estimates in the optimal segments,

TABLE IV: Reconstruction SNR for chirp and harmonic signal using either optimal segmentation or a fixed segment length matching the mean segment length of the optimal segmentation, in this case $\bar{N}_{\text{HM}} = 140$ (17.5 ms) and $\bar{N}_{\text{HCM}} = 188$ (23.5 ms). The input SNR is 10 dB.

	chirp	harmonic
opt. segm.	12.49	12.38
fixed	10.88	11.29

TABLE V: Average segment length, $\bar{N}$, for chirp and harmonic signal for different noise types at 10 dB.

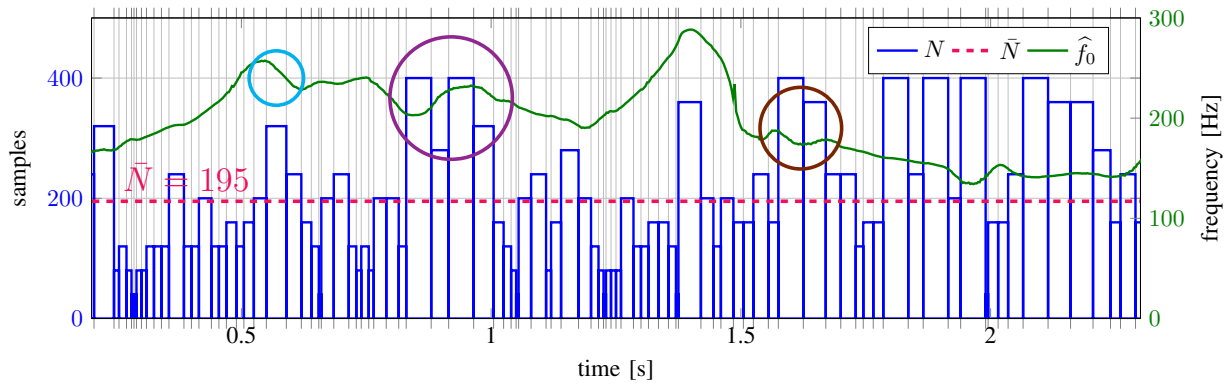
	chirp	harmonic
babble	69 (8.6 ms)	62 (7.7 ms)
FIR	73 (9.1 ms)	65 (8.1 ms)
LP	119 (14.9 ms)	91 (11.4 ms)

meaning that in some cases 40 samples (5 ms) are reconstructed based on one estimate of fundamental frequency and chirp rate, whereas in other cases, 400 samples (50 ms) are estimated based on one estimate. This is compared to estimates from segments with a fixed length where the length of the segments is set to the mean length of the segments from the optimal segmentation. In this case, $\bar{N}_{\text{HM}} = 140$ (17.5 ms) and $\bar{N}_{\text{HCM}} = 188$ (23.5 ms). This means that the reconstructions based on optimal segmentation and fixed segment length use the same number of segments to represent the signal. The reconstruction SNR is shown in Table IV. The table shows that with the same number of segments used for the reconstruction, a better reconstruction SNR can be obtained when optimal segmentation is used instead of using a fixed segment length. The reconstruction SNR is more than 1.5 dB better for the chirp model and more than 1 dB better for the traditional harmonic model when comparing optimal segmentation to a fixed segment length. By comparing the harmonic chirp model to the traditional harmonic model, a better reconstruction SNR is obtained with the harmonic chirp model when optimal segmentation is used, even though the chirp model uses only 109 segments and the traditional harmonic model uses 147 segments to represent the entire signal.

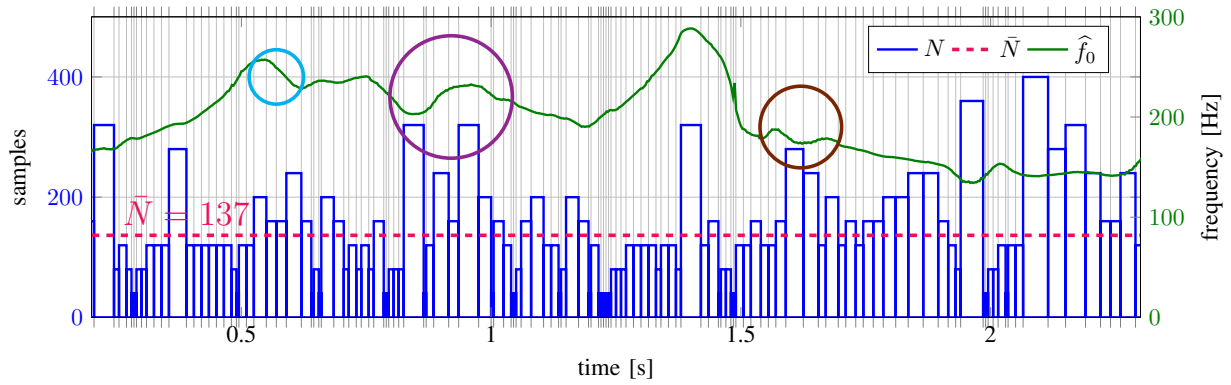
The segmentation is also tested for the signal in babble noise and prewhitened babble noise at an input SNR of 10 dB. The average segment lengths in the different cases are shown for the two models in Table V. In all cases, the signal is divided into longest segments when the chirp model is used. With respect to the different noise scenarios, the tendency is the same for the two models. The segments are shortest when the signal in babble noise is considered, followed closely by the prewhitened signal using FIR filtering. The longest segments are obtained with the LP filtered signal.

VIII. Conclusion

Traditionally, non-stationarity, fixed segment lengths and noise assumptions have limited the performance of fundamental frequency estimators. In this paper, we take these factors



(a) Harmonic chirp model



(b) Traditional harmonic model

Fig. 14: Segment length as a function of time for (a) the harmonic chirp model and (b) the traditional harmonic model. The average segment length, $\bar{N}$, is marked by the red line. The average segment length is 195 samples (24.4 ms) for the harmonic chirp model and 137 samples (17.1 ms) for the traditional harmonic model. The total number of segments is 105 for the chirp model and 150 for the harmonic model.

into account. We described the voiced part of a speech signal using a harmonic chirp model that allows the fundamental frequency to vary linearly within each segment. We proposed an iterative maximum likelihood estimator of the fundamental frequency and chirp rate based on this model. The estimator reaches the Cramer-Rao lower bound and shows expected correspondence between the estimate of the fundamental frequency and fundamental chirp rate of speech. Based on the maximum a posteriori (MAP) model selection criterion, the chirp model was shown to be preferred over the traditional harmonic model for long segments, suggesting that the chirp model is better at describing the non-stationary behaviour of voiced speech. Since the extent of the non-stationarity of speech changes over time, a fixed segment length is not optimal. Therefore, we also proposed varying the segment length based on the MAP criterion. Longer segments were obtained when the chirp model was used compared to the traditional harmonic model, again suggesting a better fit of the model to the speech. The maximum likelihood and MAP estimators are based on an assumption of white Gaussian noise. However, in real life the noise is rarely white. Therefore, we also suggested using two filters to prewhiten the noise, a simple FIR filter and one based on linear prediction (LP). They both have a minor influence on the speech signal, but the LP filter gives less error on the fundamental frequency estimate when the noise level is high. Further, the LP filter gives longer segment lengths in the optimal segmentation.

REFERENCES

[1] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2014, pp. 1400–1404.

[2] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80, no. 9, p. 19371944, 2000.

[4] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 409–412.

[5] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004.

[6] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering." *EURASIP J. on Advances in Signal Processing*, vol. 2011, p. 13, 2011.

[7] P. Jain and R. B. Pachori, "Event-based method for instantaneous fundamental frequency estimation from voiced speech based on eigenvalue decomposition of the hankel matrix," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 10, pp. 1467–1482, 2014.

[8] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[9] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint doa and pitch estimation," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[10] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *J. Acoust. Soc. Am.*, vol. 36, no. 296, 1964.

[11] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[12] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, 1976.

[13] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

[14] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, 1993.

[15] K. Molla, K. Hirose, N. Minematsu, and K. Hasan, "Voiced/unvoiced detection of speech signals using empirical mode decomposition model," in *Int. Conf. Information and Communication Technology*, March 2007, pp. 311–314.

[16] F. R. Drepper, "A two-level drive-response model of non-stationary speech signals," *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

[17] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2013, pp. 6797–6801.

[18] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 614–624, 2009.

[19] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

[20] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[21] T. Nilsson, S. I. Adalbjornsson, N. R. Butt, and A. Jakobsson, "Multi-pitch estimation of inharmonic signals," in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[22] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech and Language Process. (TASLP)*, vol. 22, no. 12, pp. 1931–1940, 2014.

[23] P. M. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, 1996.

[24] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 3985–3988.

[25] Y. Doweck, A. Amar, and I. Cohen, "Joint model order selection and parameter estimation of chirps with harmonic components," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1765–1778, Apr. 2015.

[26] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement of non-stationary speech using harmonic chirp filters," in *Proc. Interspeech*, Sep. 2015, accepted for publication.

[27] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 2029–2032.

[28] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 646–655, 2000.

[29] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 2007, no. 1, p. 24, Jun. 2007, article ID 092953.

[30] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[31] K. V. Sørensen and S. V. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," *EURASIP J. on Advances in Signal Processing*, vol. 2005, no. 18, pp. 2954–2964, 2005.

[32] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[33] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[34] S. L. Marple, Jr., "Computing the discrete-time 'analytic' signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[35] A. Antoniou and W. S. Lu, *Practical Optimization - Algorithms and Engineering Applications*. Springer Science+Business Media, 2007.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Inc., 1993.

[37] P. M. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2118–2126, 1990.

[38] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006, vol. 1.

[39] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, 1998.

[40] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[41] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[42] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, Inc., 1998.

[43] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[44] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[45] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Process.*, Oct 2000.

[46] P. C. Hansen and S. H. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, 2005.

[47] N. S. Jayant and P. Noll, *Digital coding of wafeforms*. Prentice-Hall, 1984.

[48] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.